Constructing Genetic Maps By Rapid Chain Delineation

R.W. Doerge[1] and B.S. Weir[2]

BU-1243-M June 1994

[1] Biometrics Unit, Cornell University, Ithaca, NY 14853

[2] Program in Statistical Genetics, Department of Statistics,
North Carolina State University, Raleigh, NC 27695-8203

Running Title:

A Preliminary Method for Ordering Genetic Markers

Corresponding author:

R.W. Doerge

324 Warren Hall

Biometrics Unit

Cornell University

Ithaca, NY 14853

Phone number: 607-255-2461

Abstract

The construction of genetic maps from pairwise recombination data is considered. An intuitive algorithm for gaining both linkage groups and order within linkage groups is presented as a rapid chain delineation process (RCD). Genetic ordering problems are analogous to the historic *traveling salesman problem* in which a salesman is asked to travel between cities in the shortest possible route. The proposed approach works for an arbitrarily large number of progeny scored at an arbitrary number of genetic markers, while remaining computationally simple. Comparative application to Buetow and Charkravarti's seriation method, as well as MAPMAKER is presented, demonstrating the benefits of RCD's simplicity and speed. Simulated maps are presented for both evenly and unevenly spaced markers for the purpose of assessing the performance of RCD over increasing sample size, marker number, and total map distance per linkage group. Lastly, published genetic data from human chromosome 21 and chromosome 22 are used to construct respective genetic maps using RCD, and are compared to simulated annealing.

Key Words:

pairwise recombination data, framework maps, linkage groups, objective functions

Introduction

Over the past decade there has been an increasing amount of attention paid to the problems of locating quantitative trait loci (QTL), the genes responsible for quantitative traits. While these issues are certainly important, we point out that there exists an underlying complexity to the search for QTL which begins with division of genetic markers into linkage groups, for the eventual representation of a specific regions of the genome. The initial grouping is based upon results of pairwise comparisons of recombination fractions between all available markers (or loci). Within each linkage group, markers are ordered to provide a structural foundation for the search of QTL, without which locating QTL, relative to the entire genome, is essentially impossible. It is to this end one should take into account a maximal amount of information available from the data while at the same time making a minimum number of calculations.

The need for ordering and placing large numbers of markers on genetic maps offers computational challenges as markers are being scored in species without framework maps. Our colleagues in forestry, for example, can generate data for several hundred RAPD markers on megagametophytes from a single tree in the space of a few weeks (Grattapaglia et al., 1992), and propose to repeat this exercise on many different trees, each having different segregating markers. Their need is for a routine method of defining linkage groups, and obtaining an initial ordering of genetic markers within linkage groups when the number of group members is too large for an exhaustive search over all orders, while at the same time requiring little user interaction. With this as our motivation, we note that the closer two loci are to each other on a chromosome the more likely the genes at those two loci will be transmitted together to the progeny. It is desired to order the m loci so that the order which is closest to the true order may be achieved. Ideally, we would like to consider all possible orders of the m loci, with the maximum overall likelihood providing the best order, but as the number of loci increases the $m!/2$ possible orderings become too numerous to investigate. For example, thirty markers have over 1×10^{32} possible orderings. It becomes

necessary to develop a method of ordering markers which will produce an optimal map with the highest likelihood of occurring, when compared to any other arrangement, while not requiring the complete enumeration of all possible orders.

Historically, linkage groups and ordering markers within linkage groups are established by multipoint analysis (Lathrop et al., 1985). Extensive research has been done in the area of human linkage analysis, and although the research presented does not apply solely to human linkage analysis, it is upon this knowledge that the following discussion is based. This problem is not new, and several useful methods have been described, including branch and bound methods (Thompson, 1984), simulated annealing (Corana et al., 1987; Weeks and Lange, 1987; Falk, 1992), and seriation (Buetow and Chakravarti, 1987a,b). The seriation method comes closest to meeting our goals, and in this paper we offer an alternative strategy which has the advantage of computational simplicity, as well as speed. We are encouraged in our presentation of the Rapid Chain Delineation (RCD) method by some comments made by Thompson (1984):

> A heuristic scoring criterion for estimation or inference is never as satisfactory as an estimate based on likelihood, but the greater ease of computing such a heuristic estimate may lead to its widespread use. The statistical justification for such a method must lie in the accuracy of results and/or in the approximation of such results to those obtained via some valid method of inference, such as by maximum likelihood.

Methods

Ordering Markers: Rapid Chain Delineation Algorithm

This paper proposes a simple method which begins with a two-point analysis, and ends with a preliminary genetic map. Each marker is first checked to assure independent segregation, and then a criterion based upon a recombination value cutoff is used to form linkage groups. Order within

each of the established linkage groups is formed by the Rapid Chain Delineation (RCD) algorithm. RCD can effectively handle mating schemes for backcross, F_2 dominant, F_2 codominant, and RI models, or accept previously estimated recombination fractions as the data from an model.

It is necessary to check for independent segregation since systematic segregation distortion will cause markers to segregate in an irregular fashion (i.e. not 1:1), and incorrect estimates of linkage will reflect this distortion. Independent segregation of all single genetic markers is verified in order to gain an idea of the true linkage between pairs of markers. Markers which are segregating independently are considered in pairs with a cutoff criterion, based on either a chi-square or recombination value cutoff, for the purpose of grouping markers into linkage groups, while at the same time eliminating some of the $m!/2$ possible orderings. The cutoff criterion when based upon a chi-square test statistic for markers i and j is a function of sample size and estimated recombination fraction

$$\begin{aligned} X_{ij}^2 &= \frac{(n-2k)^2}{n} \\ &= n(1-2\hat{r})^2, \end{aligned} \tag{1}$$

where k is the number of recombinants, and n is the total number of individuals scored. The appropriate rejection region is related to the estimated recombination value $\hat{r}$ by the following relation (Doerge, 1993a,b)

$$\Pr\left[\frac{(n-2k)^2}{n} \leq \chi_1^2\right] = \Pr\left[\frac{\hat{r}-\frac{1}{2}}{\sqrt{\frac{1}{4n}}} < -\sqrt{\chi_1^2}\right] + \Pr\left[\frac{\hat{r}-\frac{1}{2}}{\sqrt{\frac{1}{4n}}} \geq \sqrt{\chi_1^2}\right]$$

As an example suppose we have 60 individuals scored at any number of markers, and we want to construct a genetic map based upon a maximum estimated recombination fraction $r = 0.25$. Linkage groups may be formed using the following rule: when the estimated recombination value between two markers is less than $r = 0.25$, the markers are considered to be linked. Likewise, the corresponding chi-square cutoff (Eq. 1) for $r = 0.25$ is 15, and has a significance level of $2\alpha = 0.000215$ (Doerge, 1993b). A linkage group is formally defined as a set of markers in which

each marker is linked with at least one other marker in the same set. Under this definition, not all pairs of markers in a linkage group have to satisfy the specified cutoff criterion of the RCD method. The definition merely states that a marker in a linkage group has to be linked to at least one other marker in that linkage group by at least the cutoff criterion.

RCD: First Stage

Once all markers have been checked for independent segregation, linkage groups are formed. Linkage between markers i and j is tested using the following hypotheses, based upon their estimated pairwise recombination fraction $\hat{r_{ij}}$,

$$H_0 : r_{ij} = 0.5$$

$$H_a : r_{ij} < 0.5.$$

A linkage group is defined to consist of those loci for which this null hypothesis is rejected at least once between that locus and other loci in the group. For large numbers of markers, stringent rejection criteria may be needed to avoid large linkage groups. Raising the significance level α increases the total number of markers added to the linkage groups, while at the same time decreases the number of linkage groups.

After the initial cutoff criterion of the RCD method is satisfied, the resulting markers make up the working marker pool, and their corresponding estimated recombination values are sorted in ascending magnitude. A chain is created from the marker pool beginning with the two markers which have the smallest estimated recombination fraction. Additional markers are added to the chain and deleted from the marker pool on the basis of the preceding smallest estimated recombination fractions, and whether or not the marker already has been added to the chain. When no more markers are added to the current chain, a linkage group is declared having an initial ordering based upon the manner in which the markers were added to the linkage group. Linkage groups continue to be formed until all markers have been attempted. It is possible for a

marker to remain unlinked to any of the linkage groups.

As an example of the first stage of the RCD algorithm, consider five markers, A through E. Using a chi-square cutoff criterion 15, assume that all markers segregate independently. Suppose that data existed, and the corresponding matrix of chi-square test statistics for each pair of markers is

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} B & C & D & E \end{array} \\ \left[\begin{array}{cccc} 35.9 & 0.01 & 5.00 & 12.0 \\ & 7.30 & 11.8 & 10.1 \\ & & 20.7 & 12.5 \\ & & & 19.4 \end{array}\right] \end{array}$$

and the estimated two point recombination fractions for each pair of markers are

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} B & C & D & E \end{array} \\ \left[\begin{array}{cccc} 0.09 & 0.56 & 0.50 & 0.55 \\ & 0.50 & 0.51 & 0.50 \\ & & 0.17 & 0.30 \\ & & & 0.16 \end{array}\right] \end{array}$$

Starting with arbitrary marker A, the only chi-square statistic greater than 15 is that between marker A and marker B. Neither A nor B are linked to any other marker, so $A - B$ defines one linkage group. The order of the markers within this linkage group is irrelevant since there are only two markers. Now A and B are eliminated from the marker pool, and the process starts over with marker D (or any other remaining marker). Marker D is linked to marker E ($\chi^2 = 19.4$) and marker D is appears linked to marker C ($\chi^2 = 20.7$). Note that marker C and E are appear not linked ($\chi^2 = 12.5$) to each other, but are in the same linkage group, thus satisfying the definition of a linkage group. Since no more markers can be added to the D, C, E linkage group, because the marker pool has been exhausted, it remains to order these markers based upon estimated recombination fractions. Beginning with marker D, marker E has the smallest recombination (0.16), so markers $D - E$ initiate the chain. Marker C has the smallest recombination with marker D (0.17) so the final order of the second linkage group is $C - D - E$.

Objective Functions

As just shown, the first stage of the RCD method defines linkage groups and initial orders simultaneously. As a means of comparing orders, we define an *objective function.* Assuming independence of recombination fractions, we consider the following two objective functions: 'the sum of adjacent recombinations' (SAR), and the 'sum of adjacent log-likelihoods' (SAL). The SAR is the sum of recombination fractions over the given order, and although recombination fractions do not add linearly, the SAR does provide a means by which orders may be compared. The lowest SAR may be used as an indicator of the best order found. As a result of the assumption of independence, we can also use the sum of adjacent log-likelihoods (SAL) over all intervals in an order, as the objective function. Using a likelihood function based on the mating scheme at hand, we simply employ the estimated recombination fraction between markers i and j in the appropriate likelihood function, along with the necessary genotypic counts. The objective function for a given order is

$$\text{SAL} = \sum_{i=1}^{m-1} \log\left[L_i(r)\right],$$

where $i = 1, ..., m-1$ defines the $m-1$ intervals between the m ordered markers in a linkage group. The maximum SAL may be used to indicate the best order. This objective function makes it possible to provide the relative improvement (i.e. how many times more likely) of the maximized likelihood order over the next best order by calculating the difference of the two SALs, and raising ten to the power of this difference. For the purpose of the RCD method, the user decides which objective function to use in the analysis.

RCD: Second Stage

The second stage of the RCD algorithm is a continuation of the initial order in that systematic modifications of the existing order eventually provides the final order. Order modifications are assessed via their objective functions.

The first level of modification is a switching of each of the $m(m-1)/2$ pairs of loci in turn. If the switch produces a better order, i.e. a lower SAR or higher SAL, the new order is adopted, and the process begins again with the first marker in the chain. This pair switching terminates when no complete set of switches produces an improvement in the objective function. The second level of modification is based on all $(m-2)$ successive triplets of adjacent loci. All three permutations for each triple are assessed via the respective objective function, and the one contributing the best objective function score is accepted. As with the first level of modification, once a new order has been adopted, the process begins again with the first triple in the new order. If the current triple provides no improvement in the objective function, the next triple along the chain, which includes the last two members of the previous triple is evaluated in the same manner. When all triple permutations are have been treated, and no improvement in the objective function is demonstrated, the RCD map is obtained. A third level of modification which is optional, consists of trying every marker in turn in every existing interval of the current genetic map.

Due to the manner in which the first stage of the RCD algorithm adds markers to the chain, the initial order obtained is most likely close to the optimal order obtainable by this method, since the smallest estimated recombination fractions determine chain extension. However, when markers are very close together, the manner in which the proposed method initiates and adds to the chain is not optimal, necessitating the need for the second stage. As an example of this phenomenon, consider the following recombination matrix for four markers A, B, C, D, and use the SAR as the objective function for ease of calculation.

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 0.09 & 0.19 & 0.17 \\ & 0 & 0.26 & 0.22 \\ & & 0 & 0.32 \\ & & & 0 \end{bmatrix} \end{array}$$

The Rapid Chain Delineation (RCD) method begins the chain with markers A and B ($r_{AB} =$

0.09). Marker D is then added to the left of marker A since this is the next smallest recombination fraction ($r_{DA} = 0.17$). Last, marker C is added to the right of marker B ($r_{BC} = 0.26$). The initial chain order from stage one of the RCD method is $D - A - B - C$ and has an SAR of 0.52. Stage two of RCD evaluates permutations of all possible pairs of markers, and then units of three in current chain so as to eliminate the problems of the initial stage, and finds the optimal order $D - B - A - C$ which has an SAR of 0.50, the minimum SAR over all 4! orderings.

Results

Performance of RCD

The only way to assess performance of an ordering algorithm is to apply it to a set of loci of known order. In the first place we used the same simulation strategy Buetow and Chakravarti (1987b) adopted when they evaluated the performance of the seriation method. Three sets of estimated recombination values were simulated according to specified maps 5U, 10U and 5N (Figure 1), to represent uniformly and nonuniformly spaced loci. In each case, the estimated recombination fraction for loci i and$j$ was the proportion of n trials resulting in a "success" where the number of successes was binomially distributed with parameters n and r_{ij}. Sample size n was considered at values of 20, 40, 60, 80, 100. The r_{ij} values followed from the maps by use of Haldane's mapping function, and were given explicitly in Table 1 of Buetow and Chakravarti (1987b). The proportions of times in 100 replicates of this process that led to the correct order are shown in Table I, along with the corresponding values Buetow and Chakravarti (their table 3) for seriation. These latter values also show the number of 100 replicates that allowed an order to be found. We see from this simulation that the two methods are comparable, with RCD out-performing seriation in each case.

It appears that the method of seriation can be flawed by inversions of two closely linked loci.

In fact, when seriation finds an incorrect order it is usually different from the correct order by only a single inversion. The issue of incorrect orders in seriation is said (Buetow and Chakravarti, 1987) not to be due to failure of the algorithm, but rather to be due to an attempt of the method to reach an optimal order based upon a 'poor' sample. A poor sample is one in which the sample size consists of less than one hundred meioses. Of course, the correct order for the sample may not agree with the correct order for the population from which the sample is drawn.

The RCD method was also compared to the MAPMAKER/EXP program (Lander et al., 1987; Lincoln et al., 1992a,b). This package contains a test set of data, labeled f2.raw, containing 333 F_2 individuals scored at 20 RFLP loci. Not all loci were scored in every individual. Only estimated recombination values of less than 0.4 are used under the MAPMAKER/EXP protocol, so initially, the same rule was adopted to define the linkage groups in the RCD method. The analysis of the f2.raw data set in MAPMAKER/EXP provides an additional data set, labeled f2.2pt, containing the estimated two-point recombination values. Using the estimated two-point recombination fractions from MAPMAKER/EXP, RCD achieved the same genetic map (Table II). However, for different recombination value cutoffs, the genetic map changes significantly (Table II).

A third analysis of the RCD method used simulations which were conducted for the cases of 10, 20 or 30 markers, spaced uniformly between 1 and 49 centiMorgans apart, or spaced randomly in maps of the same total length as those for uniform spacing. In Figure 2 the proportions of 100 replicates that led to the correct order are plotted for each sample size considered (25, 50, 100, 250). For the parameters chosen, there is obviously a strong case for having the samples with at least 100 informative meioses. Beyond that level, there is better performance for evenly than unevenly spaced markers. The performance graphs provided in Figure 2 give an idea of the number of markers per linkage group and individuals one should consider in order to maximize the capability of the RCD method. Lastly, we can see the direct benefit toward constructing

genetic maps with the markers as uniformly spaced as possible, while remaining in the range of RCD performance which provides the best results.

A fourth investigation involved both simulated and real data as investigated by Falk (1992). First, the two-point recombination values as simulated for the 3rd Genetic Analysis Workshop (GAW3, MacCluer et al., 1985) were used to compare RCD to Falk's simulated annealing approach. Table III shows MDMAP (1992) and RCD obtain the lowest possible SAR. Both methods use SAR as an objective function.

Second, two human data sets from chromosome 21 (Warren et al., 1989), and chromosome 22 (Rouleau et al., 1989) were analyzed using RCD over different cutoff values, and compared to published results for chromosome 21 (Warren et al., 1989), chromosome 22 (Rouleau et al., 1989), and Falk (1992). Table IV and Table V summarize the outcome of this comparison, using SAR as the objective function. RCD performed better than the published results of Warren et al. (1989) and Rouleau et al. (1989), and as well as the published results for Falk's simulated annealing approach (1992). However, since RCD allows a cutoff criterion to define the construction of the genetic map, limiting the recombination value allowed in the map restricts the use of some genetic markers. Therefore, (Table IV and Table V) a lower cutoff criterion for a dense framework map, may eliminate one or two markers, and achieve a lower SAR.

Finally, for the purpose of evaluating the performance of RCD with missing data, simulations were conducted for samples of size 50, 100, 250 using 20 loci. Figure 3 illustrates the performance of the RCD method as the percent data missing approaches 50%. As we would expect, for large sample size (250), the performance of RCD becomes affected when more than 20% of the data is missing.

Discussion

The Rapid Chain Delineation (RCD) algorithm leads to ordered linkage groups as part of mapping projects, and as preliminary to searches for genes affecting quantitative traits. There is no guarantee that the RCD order is the best possible order, but probabilities upwards of 90% of obtaining the correct order from samples of 100 or more meioses, and adjacent markers not more than 20 cM apart seem quite feasible. The real advantage of RCD seems to arise through its speed and ease of use. The method can be programmed to work noninteractively, and was able to construct a map of 210 RAPD markers in 18 linkage groups using a recombination values cutoff of $r = .30$, and the SAR as the objective function in less than 10 seconds on a SPARC-2 workstation. RCD was not compared to the speed of Falk's simulated annealing approach because the parameters required by simulated annealing were not provided (Falk, 1992). It is well known that the initial parameters used to implement simulated annealing greatly affect its speed.

The preliminary order gained from RCD may be used alone as an exploratory measure, or in conjunction with other genetic map construction methods to provide a more expedient means of gaining genetic maps.

Acknowledgments

The authors thank Paul Lewis for many constructive and helpful discussions. This material is based upon work supported in part by the Program in Mathematics and Molecular Biology at the University of California at Berkeley, which is supported by the National Science Foundation under grant DMS-8720208 and in part by NIH grant GM32518. The Government has certain rights to this material.

References

Buetow KH, Chakravarti A (1987): Multipoint gene mapping using seriation. I. General methods. Am J Hum Genet 41:180-188.

Buetow KH, Chakravarti A (1987): Multipoint gene mapping using seriation. II. Analysis of simulated and empirical data. Am J Hum Genet 41:189-201.

Corana, A, Marchesi, M, Martini, C, and Ridella, S (1987): Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. ACM Trans on Math Software 13:2:262-80.

Doerge, RW (1993): Statistical Methods for Locating Quantitative Trait Loci with Molecular Markers. Ph.D. dissertation. North Carolina State University, Raleigh, NC.

Doerge, RW (1993): Testing for linkage: phase known/unknown. Journal of Heredity (submitted).

Falk CT (1992): Preliminary ordering of multiple linked loci using pairwise linkage data. Genet Epidemiol 9:367-375.

Grattapaglia D, Chaparro J, Wilcox P, McCord S, Werner D, Amerson H, McKeand S, Bridgewater F, McIntyre L, Doerge R, Weir B, Whetten R, O'Malley D, Sederoff R (1992): RAPD mapping and tree improvement. The International Conference on the Plant Genome.

Lander ES, and Green P (1987): Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci 84:2363-2367.

Lathrop G, Lalouel J, Julier C, Ott J (1985): Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. Am J Hum Genet 37:482-498.

Lincoln S, Daly M, Lander E (1992): Constructing Genetic Maps with MAPMAKER/EXP 3.0. Whitehead Institute Technical Report. 3rd edition.

Lincoln S, Daly M, Lander E (1992): Mapping Genes Controlling Quantitative Traits with MAPMAKER/QTL 1.1. Whitehead Institute Technical Report. 2nd edition.

MacCluer JW, Falk CT, Wagener DK (1985): Genetic Analysis workshop III: Multipoint mapping and linkage. AMJ Hum Genet 37:1040-1044.

Rouleau GA, Haines JL, Bazanowski A, Colella-Crowley A, Trofatter JA, Wexler NS, Conneally PM, Gusella JF (1989): A genetic linkage map of the long arm of human chromosome 22. Genomics 4:1-6.

Thompson EA (1984): Information gain in joint linkage analysis. IMA J Math Appl Med Biol 1:31-49.

Warren AC, Slaugenhaupt SA, Lewis JG, Chakravarti, A, Antonarakis SE (1989): Agenetic linkage map of 17 markers on human chromosome 21. Genomics 4:579-591.

Weeks D, Lange K (1987): Preliminary ranking procedures for multilocus ordering. Genomics 1:236-242.

TABLE I. Proportion of Correct Orders with RCD and Seriation.

No. of informative Meioses		Map		
n	Method	5U	10U	5N
20	RCD	54%	10%	68%
	Ser.*	30/97†	7/96	19/89
40	RCD	65%	31%	60%
	Ser.	63/100	30/99	42/98
60	RCD	79 %	51%	58%
	Ser.	71/100	40/100	43/100
80	RCD	91%	76%	68%
	Ser.	89/100	65/100	51/100
100	RCD	94%	85%	69%
	Ser/	93/100	80/100	54/100

*from Buetow and Chakravarti (1987)
†No. of correct orders/No. orders derived

TABLE II. Maps found by RCD and MAPMAKER/EXP for MAPMAKER f2.raw data set using MAPMAKER/EXP generated two-point recombination fractions (Lander et al., 1987; Lincoln et al., 1992a,b). RCD_k; $k = .2, .3, .4, .5$ denotes the cutoff value used to declare linkage groups, $MAPMAKER_{.4}$ is used throughout.

Method	Group	Order	SAR
$RCD_{.4}$	1	1 – 3 – 2 – 7 – 8 – 11 – 12 – 15	1.223
MAPMAKER	1	1 – 3 – 2 – 7 – 8 – 11 – 12 – 15	1.223
$RCD_{.4}$	2	9 – 10 – 16 – 4 – 5 – 6 – 17 – 18	0.824
MAPMAKER	2	9 – 10 – 16 – 4 – 5 – 6 – 17 – 18	0.824
$RCD_{.4}$	3	14 – 13 – 20 – 19	0.634
MAPMAKER	3	14 – 13 – 20 – 19	0.634
$RCD_{.5}$	1	9-10-16-4-5-6-17-18-12-15-11-8-7-2-3-1-14-13-20-19	3.637
$RCD_{.3}$*	1	9-10-16-4-5-6-17-18	0.824
	2	15-12-11-8-7-2-3-1	1.223
	3	13-20-14	0.256
$RCD_{.2}$†	1	4-5-6-17-18	0.391
	2	9-10-16	0.208
	3	2-3-1	0.188
	4	13-20-14	0.256
	5	7-8-11	0.285

*Marker 19 not placed.
†Markers 12,15,19 not placed.

TABLE III. RCD compared to Falk's simulated annealing approach MDMAP (1992), using data for four linked loci as simulated for the 3rd Genetic Analysis Workshop (GAW3, MacCluer et al., 1985). All possible orders of are also given, showing that MDMAP and RCD both achieved the lowest SAR.

Method	Order	SAR
MDMAP	B–H–F–C	0.331
RCD	B–H–F–C	0.331
	All Possible Orders	
	B-H-F-C	0.331
	H-B-F-C	0.346
	F-B-H-C	0.376
	C-B-H-F	0.378
	B-F-H-C	0.386
	C-B-F-H	0.403
	B-H-C-F	0.522
	H-B-C-F	0.539
	B-F-C-H	0.547
	B-C-F-H	0.549
	B-C-H-F	0.579
	F-B-C-H	0.594

TABLE IV. Markers from Chromosome 21 as ordered by MDMAP (Falk, 1992), Warren et al. (1989), and RCD_k, where $k = .1, .2, .3, .4, .5$ is the range of cutoff values for linkage groups.

Method	Order	SAR
MDMAP Genetic Epidemiology 9:371	1-16-2-5-13-3-14-10-9-11-4-8-7-15-6-12	0.96
Warren et al. Genomics 4:579	1-16-2-5-13-3-10-14-9-11-4-8-15-7-12-6	1.39
$\text{RCD}_{.5}$	1-16-2-13-5-3-14-10-9-4-11-12-6-15-7-8	1.13
$\text{RCD}_{.4}$	1-16-2-5-13-3-14-10-9-11-4-8-7-15-6-12	0.96
$\text{RCD}_{.3}$	1-16-2-5-13-3-14-10-9-11-4-8-7-15-6-12	0.96
$\text{RCD}_{.2}$*	1-16-2-13-5-3-14-10-9-11-4-8-15-6-12	0.94
$\text{RCD}_{.1}$†		
Group 1	5-13-3-14-10-9-11-4	0.24
Group 2	7-15-6	0.02
Group 3	1-16-2	0.14

*Marker 7 not placed.
†Markers 8 and 12 not placed.

TABLE V. Markers from Chromosome 22 as ordered by MDMAP (Falk, 1992), Rouleau et al. (1989), and RCD_k, where $k = .1, .2, .3, .4, .5$ is the range of cutoff values for linkage groups.

Method	Order	SAR
MDMAP Genetic Epidemiology 9:373	2-1-3-4-5-6-7-10-8-9-11-12-14-13	0.85
Rouleau et al. Genomics 4:579	1-2-3-4-5-6-7-8-9-10-11-12-13-14	1.17
$RCD_{.5}$	2-1-3-4-5-6-7-10-8-9-11-12-14-13	0.85
$RCD_{.4}$	2-1-3-4-5-6-7-10-8-9-11-12-14-13	0.85
$RCD_{.3}$	2-1-3-4-5-6-7-10-8-9-11-12-14-13	0.85
$RCD_{.2}$	2-1-3-4-5-6-7-10-8-9-11-12-14-13	0.85
$RCD_{.1}$*		
Group 1	2-1-3-4-5-6-7-8-9	0.35
Group 2	13-14-12	0.10

*Markers 10 and 11 not placed.

Figure 1

5U $A \overset{10}{--} B \overset{10}{--} C \overset{10}{--} D \overset{10}{--} E$

10U $A \overset{10}{--} B \overset{10}{--} C \overset{10}{--} D \overset{10}{--} E \overset{10}{--} F \overset{10}{--} G \overset{10}{--} H \overset{10}{--} I \overset{10}{--} J$

5N $A \overset{5}{--} B \overset{1}{--} C \overset{9}{--} D \overset{16}{--} E$

Figure 2

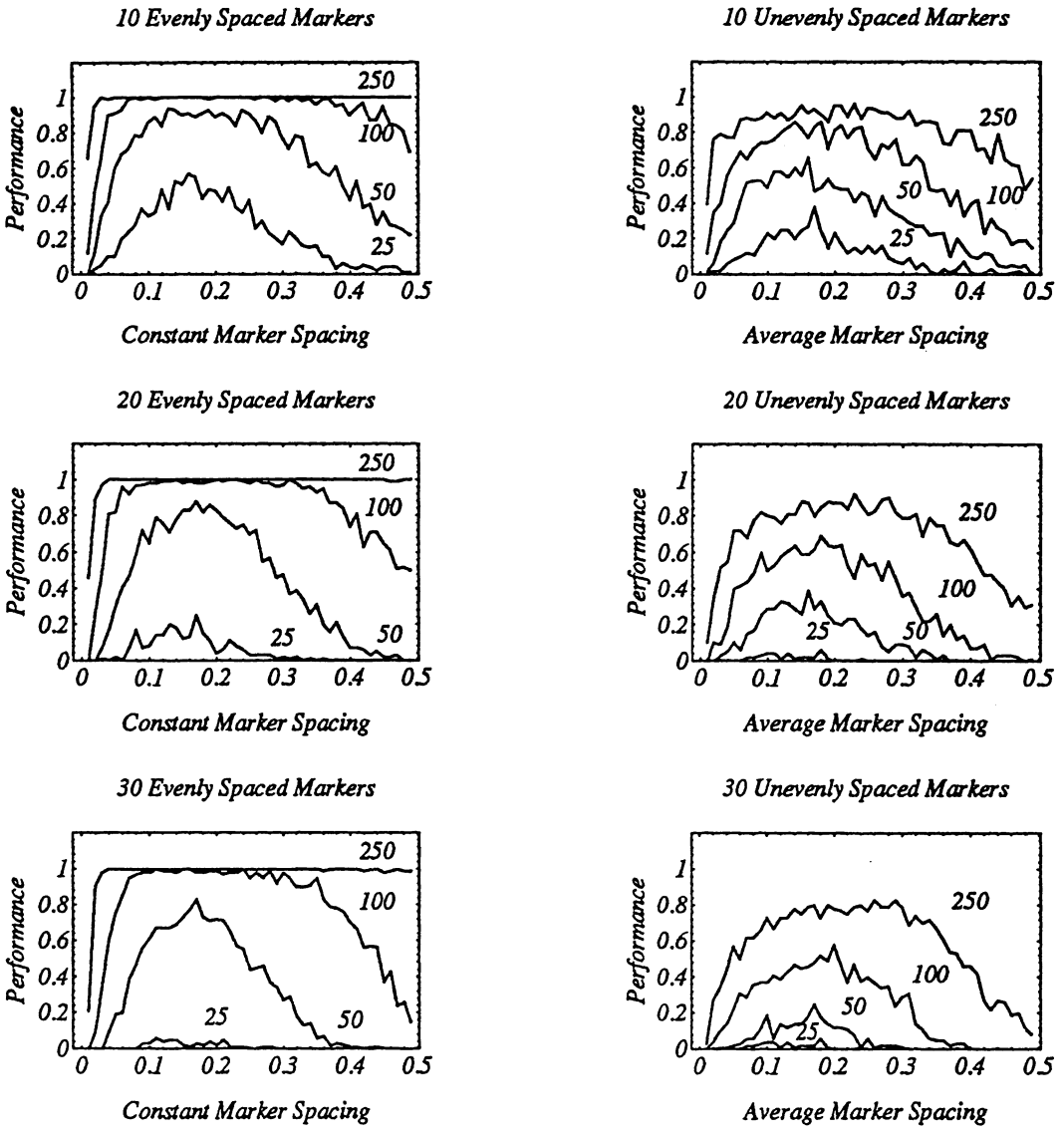
10 Evenly Spaced Markers
Performance
Constant Marker Spacing
250
100
50
25
10 Unevenly Spaced Markers
Average Marker Spacing
20 Evenly Spaced Markers
20 Unevenly Spaced Markers
30 Evenly Spaced Markers
30 Unevenly Spaced Markers

Figure 3

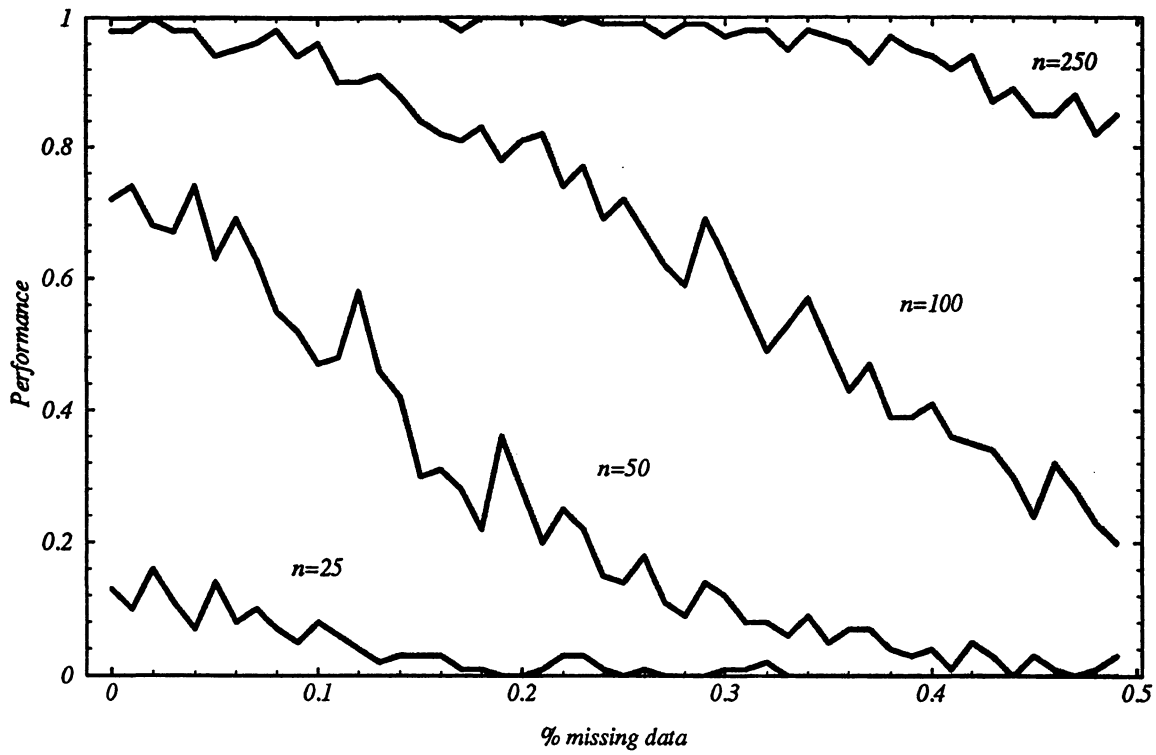
1
0.8
0.6
0.4
0.2
0
Performance
0
0.1
0.2
0.3
0.4
0.5
% missing data
n=250
n=100
n=50
n=25

Figure 1: Maps used in simulations for comparisons with seriation. Distances are in centiMorgans. 5U and 10U represent 5 and 10 uniformly spaced markers, while 5N represents 5 unevenly spaced markers.

Figure 2: Performance of RCD for evenly and unevenly spaced markers (10, 20, 30) over increasing sample size (n=25, 50, 100, 250), and increasing distance between adjacent markers.

Figure 3: Performance of RCD for 20 markers distributed uniformly over a $200cM$ map with increasing sample size (n=50, 100, 250), and increasing percentage of data missing.